

1 **SciRide Finder : a citation-based paradigm in biomedical literature search**

2 Adam Volanakis¹, Konrad Krawczyk^{1*}

3 ¹ SciRide.org

4 * Corresponding Author

5 **Abstract**

6 There are more than 26 million peer-reviewed biomedical research items according to
7 Medline/PubMed. This breadth of information is indicative of the progress in
8 biomedical sciences on one hand, but an overload for scientists performing literature
9 searches on the other. A major portion of scientific literature search is to find
10 statements, numbers and protocols that can be cited to build an evidence-based
11 narrative for a new manuscript. Because science builds on prior knowledge, such
12 information has likely been written out and cited in an older manuscript. Thus, Cited
13 Statements, pieces of text from scientific literature supported by citing other peer-
14 reviewed publications, carry significant amount of condensed information on prior
15 art. Based on this principle, we propose a literature search service, SciRide Finder
16 (finder.sciride.org), which constrains the search corpus to such Cited Statements only.
17 We demonstrate that Cited Statements can carry different information to this found in
18 titles/abstracts and full text, giving access to alternative literature search results than
19 traditional search engines. We further show how presenting search results as a list of
20 Cited Statements allows researchers to easily find information to build an evidence-
21 based narrative for their own manuscripts.

22

23 **1. Introduction**

24

25 More than 60,000 articles are deposited in PubMed each month, making literature
26 search an increasingly difficult task¹. A typical literature query consists of keyword-
27 based search by services such as Google Scholar, PubMed, Scopus or Web of
28 Science²⁻⁴. The results typically consist of a list of titles and abstracts from documents
29 that contain the query keywords. The scientist is then tasked with parsing through an
30 extensive list of results, to extract information directly from titles/abstracts or to

31 follow a link to the full document.

32

33 As such literature search can be burdensome, intelligent text mining of scientific
34 publications has been seen as an alternative for extracting and organizing information
35 from the ever-growing PubMed collection⁵. Sites such as iHOP or Chilibot mine
36 field-specific knowledge by collating information regarding biomolecules from
37 millions of PubMed publications^{6,7}. Less field-specific services such as COLIL,
38 provide a service showing comments in more recent research on older manuscripts⁸.
39 These tools demonstrate that strategic text mining and intelligent filtering can lead to
40 new, more efficient tools for biomedical literature search.

41

42 Strategic text mining can be used to separate relevant information from tangential
43 text. For instance, because of legal restrictions, typical literature search engines
44 operate on the remit of copyright-available titles and abstracts alone, whereas full text
45 contains more pertinent information⁹. For instance, tools such as Biotext or Yale
46 Image Finder allow searches in Figure or Table captions alone in order to identify
47 relevant information only^{10,11}. To understand what information is potentially
48 irrelevant, it is necessary to identify portions of searchable documents that can be of
49 more interest to the person performing literature search.

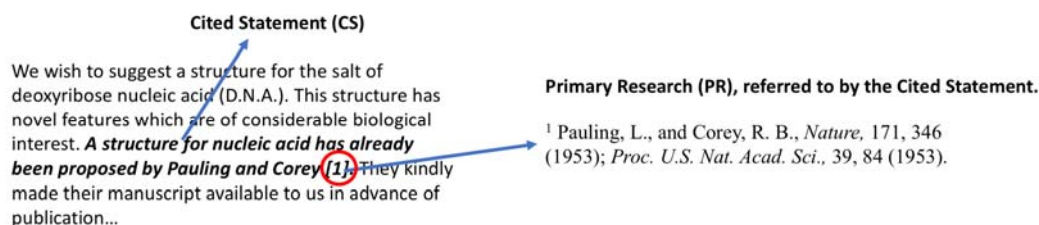
50

51 One major aim of literature search is to identify earlier papers to support the narrative
52 presented in a new manuscript being written. Such narrative is constructed by citing
53 findings, numbers, data and techniques from previous publications¹². Such pieces of
54 text are easily identifiable in scientific manuscripts since they are annotated with
55 references to prior peer-reviewed publications which support the statement being
56 made^{12,13}. Therefore such statements in publications on previous literature, which we
57 here call Cited Statements, offer succinct comments on prior art, whose information
58 content is powerful enough to be used for article summarization¹⁴.

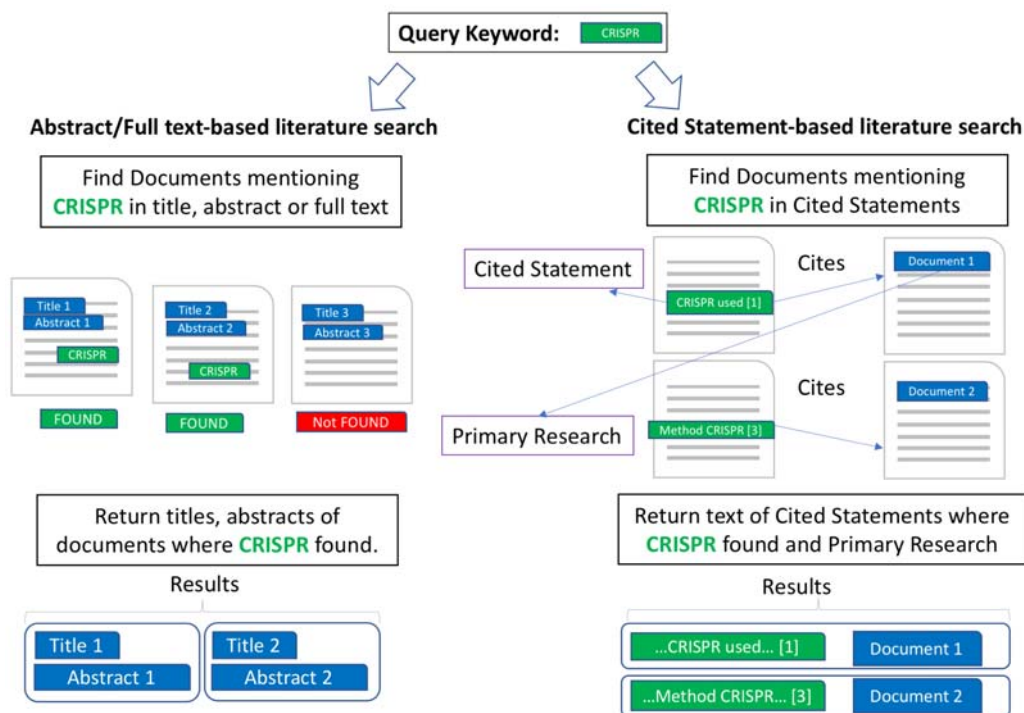
59

60 Here, we propose a simple strategy of improving text mining and literature search by
61 creating a biomedical search corpus, which is constrained to such Cited Statements
62 only. We show that Cited Statements can carry different information-retrieval data to
63 these found in titles, abstracts and full text of documents they refer to, demonstrating
64 that this methodology does not simply recapitulate information currently available in

65 scientific search engines. Furthermore, we show how presenting results in the form of
66 Cited Statement text, can offer easy access to information in several literature-search
67 scenarios. We hope that our service, available at finder.sciride.org will offer a
68 streamlined way for biomedical scientists to build evidence-based narratives for their
69 own manuscripts.



70
71 **Figure 1.** Example of a Cited Statement (CS) and Primary Research (PR). The CS is
72 shown in bold in the excerpt on the left. PR which is referred to by the CS, is shown
73 on the right. The text in the image was taken from the seminal paper by Watson and
74 Crick in 1953, entitled 'A Structure of Deoxyribose Nucleic Acid'.
75



76
77 **Figure 2.** Contrasting the traditional literature search and Cited Statement-based
78 literature search. Traditional literature search systems identify documents to be
79 retrieved by keyword hits within them and present titles and abstracts as results (left).

80 Cited Statement-based search identifies Primary Research documents by text from
81 other publications and presents the citing text.

82

83 **2. Results/Applications**

84

85 **2.1 SciRide Finder as an alternative biomedical literature search platform.**

86

87 SciRide Finder offers an orthogonal literature search strategy to platforms such as
88 PubMed or Google Scholar by focusing on Cited Statements only. In this manuscript,
89 we refer to Cited Statement (CS) as any sentence from a peer reviewed publication
90 containing citations to other manuscripts, which we refer to as Primary Research (PR)
91 (Figure 1).

92

93 We have extracted the CSs from all PubMed/Medline indexed documents where the
94 copyright allowed for data mining and reproduction. To the best of our knowledge,
95 the most suitable corpus for this task is the Open Access PubMed Central (OA PMC)
96 dataset. It is a collection of open access journals from PubMed/Medline in
97 standardized format. At the time of writing, there were approximately 1.7m
98 publications in the OA PMC dataset, which is 6% of a total of more than 26m
99 publications indexed in PubMed (or 15m if only citations with abstracts are to be
100 considered⁹).

101

102 The OA PMC dataset downloaded via the NCBI ftp service forms the core of our
103 dataset. Nevertheless, the ~1.7m OA PMC articles are only a subset of more than 4m
104 web-formatted documents available via PMC¹⁵. There are more than 2m articles
105 published after 1980 which are accessible via PMC ‘eyes-only’ subject to strict
106 restrictions on machine access and heterogeneous publisher copyrights. We therefore
107 extract such data manually if and only if the copyright situation is unambiguous.

108

109 We have set up a pipeline to collect data from the OA PMC and other publications in
110 the public domain where copyright allows it (see Materials and Methods). At the time
111 of writing, our data collection encompasses 1,786,322 peer-reviewed articles
112 contributing 43,326,402 CSs. We make this corpus accessible via efficient Lucene-
113 based search system as described in Materials & Methods. Here, we argue that our

114 CS-based search system is a new literature search paradigm, distinct from traditional
115 title/abstract and full text based methods.

116

117 The first major difference between traditional and CS -based search is the corpus
118 employed to identify documents. In traditional systems, documents are retrieved if the
119 query keywords are found within text of title, abstract or full text. On the other hand,
120 searching by CSs identifies PR documents indirectly by text contained in other papers
121 (Figure 2). CS offers an alternative commentary on the PR, by scientists who were
122 generally not involved in the original study. To prove this point, we demonstrate that
123 CSs can hold alternative information to titles/abstracts and full text of PR documents,
124 described in section 2.2.

125

126 The second major difference between traditional and CS-based search is the
127 presentation of results. In traditional literature search systems, results are presented as
128 titles/abstract, more seldom as full text excerpts. In contrast, CS-based search returns
129 the text which cites other documents. In this capacity, it identifies the information
130 which was used to build the evidence-based narrative for a manuscript: scientific
131 statements, numbers, data and techniques, all supported by prior publications. In
132 many scenarios, these are the pieces of text scientists look for in the first place to
133 build an evidence-based structure for their own manuscripts. To exemplify this, we
134 present possible applications of presenting results as CSs in section 2.3.

135

136 **2.2 Cited Statements can hold different information on documents to**
137 **titles/abstracts and full text.**

138

139 We argue that CS-based search offers a novel way of retrieving documents, that can
140 yield orthogonal results as compared to traditional search strategies. For this to be
141 true, CSs must offer distinct information-retrieval data on the PR that would not
142 normally be available by examining titles, abstract or even full text of PR document.

143

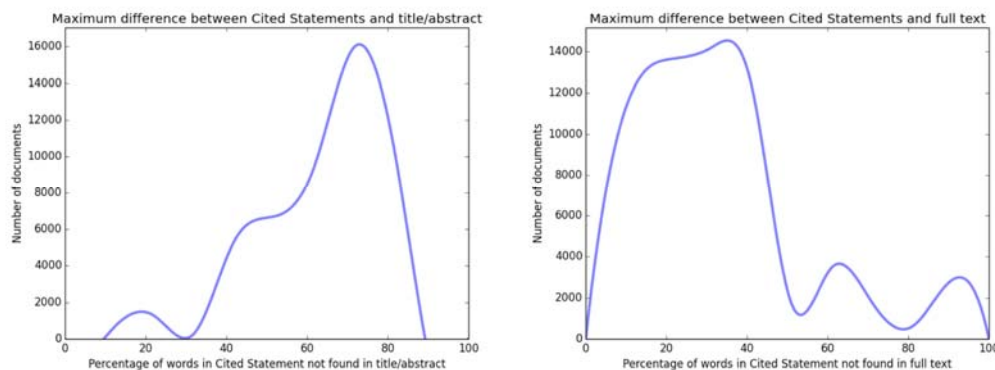
144 To quantify this, we identified 691,354 documents where we had CSs in our database
145 referring to PR documents whose full text is available for text mining. For a given CS,
146 we measured how many normalized words (stemmed, case-folded etc.) cannot be
147 found in the title/abstract and full-text of the PR documents, which we refer to as

148 'difference'. For each of 691,354 documents, we have identified the maximally
149 different (as percentage) CS with respect to the title/abstract and full text of the PR
150 document (Figure 3). These results demonstrate that for 83% of our 691,354 PR
151 documents, there exists a CS which is at least 50% different to the title/abstract of the
152 PR document. For 61% of PR documents, there exists a CS which is at least 25%
153 different to the PR document full text. Therefore, for a significant proportion of
154 publications, there exists a CS which offers information that would not be available
155 through a title/abstract or full-text search on the PR document.

156

157 We have also found that CSs tend to be different from titles/abstracts not only in the
158 extreme as above, but also on average (see Supplementary Figure 1). These results
159 demonstrate that CSs can contribute different information on the PR manuscripts than
160 title/abstract and full text. Therefore corpus constrained to CSs only can provide
161 orthogonal results to those offered by search engines which identify documents
162 directly by their titles/abstracts and full text.

163



164

165 **Figure 3.** Maximum difference between CS and text of PR. For each of 691,354
166 documents, we identify the maximally different CS with respect to title/abstract (left)
167 and full text (right) of PR. This stands to demonstrate that there are many PR
168 documents, where there exists a significantly textually different CSs referring to them.

169

170 **2.3 Applications of the Cited Statement-based literature search**

171

172 Search results from SciRide Finder do not consist of titles, abstracts or full-text
173 excerpts as is typically the case in other services, such as PubMed or Google Scholar.

174 Instead, in response to a query, we present a list of CSs, PR documents and papers
175 where the information was found (Figure 4). To exemplify the utility of presenting
176 results in this way we demonstrate possible literature search scenarios where CSs
177 instantly provide information being sought after:

178

179 *Identifying citations supporting general knowledge.* It is often problematic to identify
180 citations supporting a well-known fact. For instance, the controversy surrounding the
181 link between vaccines and autism is well known, but identifying studies discrediting it
182 is not trivial. Searching SciRide Finder for “autism” “vaccine” and “discredited”
183 would return results that debunk the notorious 1998 publication in Lancet by
184 Wakefield and colleagues.

185

186 *Identifying datasets.* Datasets used in publications are rarely cited in titles and
187 abstracts, rather being hidden in Methods sections. As an example, searching SciRide
188 Finder for the terms “ChIP-seq” “HeLa” and “Pol II” would return the publications
189 that have used datasets of RNA Polymerase II (Pol II) Chromatin
190 Immunoprecipitation sequencing (ChIP-seq) experiments in HeLa cells but also the
191 original source of these datasets (Supplementary Figure 2), thus facilitating the
192 retrieval of the datasets.

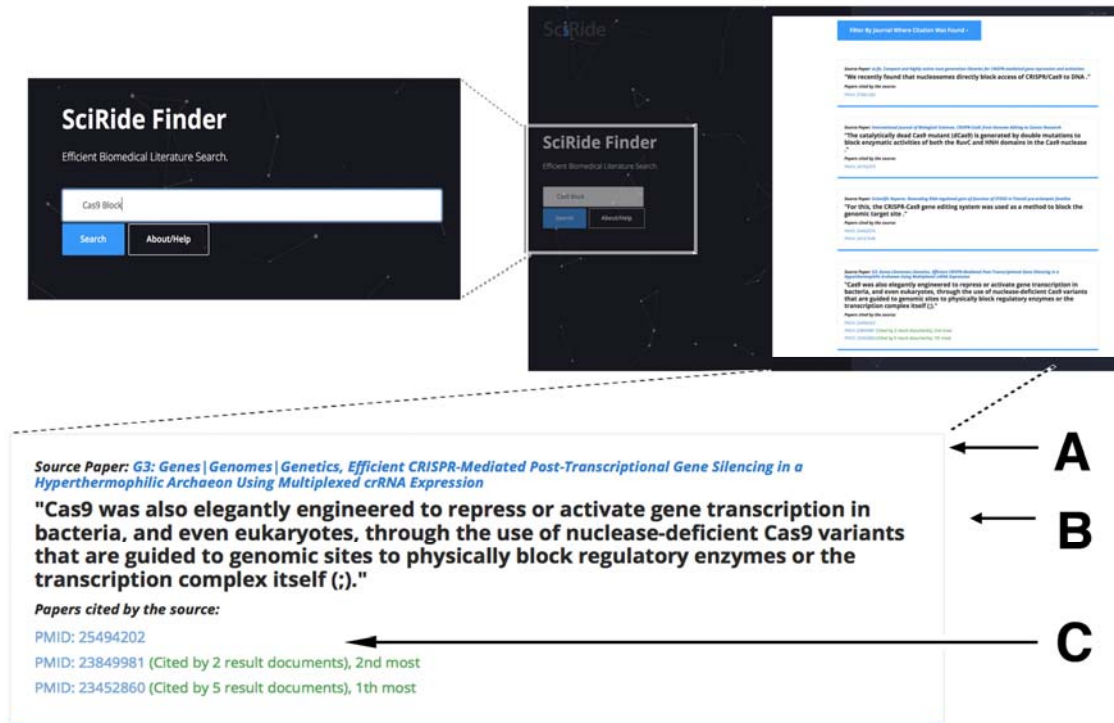
193

194 *Research technique identification.* Similarly to datasets, specific techniques used are
195 rarely available in abstracts and titles. Nevertheless, identifying publications which
196 employ a given technique or software is indispensable for reuse of protocols. For
197 example, the newly developed CRISPR/Cas9 genome editing method has an
198 alternative usage as a block for gene transcription. A PubMed search for the terms
199 “Cas9” and “Block” returns 47 publications (at the time of writing) and there is no
200 way of knowing how and in what context this method was used in each paper without
201 reading all manuscripts. The same search in SciRide Finder (Figure 4) provides a list
202 of publications where this technique was used in context. This allows us to identify
203 publications describing the method, the theory behind it, protocols used, and the
204 original research.

205

206 *Mapping connections between keywords and publications.* SciRide Finder allows for
207 searching for two or more terms appearing together, their context and the original

208 research. For example, a CS-search for ‘mRNA export’ and ‘transcription’ would
209 identify only the statements in which the two keywords appear together
210 (Supplementary Figure 3). Mapping such connections between keywords and
211 citations can be of particular interest for creating knowledge maps by the text mining
212 community.
213



226 titles/abstracts and full text. Furthermore, presenting results as CSs, is beneficial in
227 many areas of scientific literature search, whose major part is aimed at identifying
228 evidence-based pieces of text to be used in future publications.

229

230 Previous search methodologies, such as Google Scholar, aim to index all the
231 information available on documents even if the publication itself is not in the public
232 domain. On the contrary, our service indexes only a very well-defined subset of the
233 full-text articles, namely the CSs. We currently extracted ~43m CSs which contain
234 comments on 34% (or 57%, if publications without abstracts are to be omitted⁹) of all
235 of PubMed articles. This proportion should only increase as more publications
236 become open access and repositories become legally and technically unified for
237 systematic text mining¹⁵. Furthermore, since results of our service come solely from
238 open access publications, it would follow that such manuscripts would be more
239 readily cited as their content is freely accessible for scientists and search engines
240 alike¹⁶.

241

242 In summary, our system introduces an open-access, CS-only paradigm in literature
243 search. Current manifestation of this paradigm, SciRide Finder, offers an orthogonal
244 approach to reduce the burden currently associated with specific information retrieval
245 in biomedical literature. We hope that our service will facilitate the efforts of
246 researchers looking for Cited Statements, to build an evidence-based narrative for
247 their own publications.

248

249 **4. Materials & Methods**

250

251 **4.1 Data Collection for the base system – PMC Open Access Dataset.**

252

253 The OA PMC corpus was downloaded from the NCBI FTP website
254 (<ftp.ncbi.nlm.nih.gov>) and divided into sentences using Natural Language Toolkit
255 (nltk.org) and a custom set of heuristics, such as splitting text on terminal period ‘.’,
256 removing the ‘.’ from short-hands such as ‘et al.’, ‘ca.’ and normalizing the scientific
257 names (‘H. Pylori’). We identified the sentences containing citations as these having
258 the <xref> tag with attributes pointing to references section (as opposed to non-
259 bibliographic elements such as Tables and Figures). Rules were created for special

260 cases where the citation pertaining to a sentence occurs after its terminal period. Each
261 CS derived in this way contains the citing sentence, identifiers of cited articles (DOI
262 or Pubmed ID) and the metadata on the manuscript it was derived from (journal title
263 and article title). The system was set up to perform updates of this base dataset on a
264 monthly basis.

265

266 **4.2 Data Collection Beyond the PMC Open Access Dataset.**

267

268 We augment the information from the base-dataset manually from the ‘eyes-only’
269 documents where there was an unambiguous copyright situation on reproducing
270 pieces of work in a normal citation scenario. Furthermore, it is sometimes possible to
271 find the author-submitted PDF version of the document. These are documents
272 available via platforms such as BioArxiv or author homepages. Whenever we could
273 not identify a PMC version of an article, we attempted a PDF doi search. When a PDF
274 document was found in such a way, we extracted information from it using
275 PDFExtract tool from CiteSeer. PDFExtract is a utility which is capable of extracting
276 portions of a PDF-formatted scientific document and present them in machine-
277 readable plain text. Since information presented in such format is still very
278 heterogeneous, we had to create different sets of rules to interpret the PDF-extracted
279 plain text, which mostly involved detecting if the citations are number-based or
280 author-name based.

281

282

283 **4.3 Text Retrieval System.**

284

285 The Cited Statements are stored for rapid extraction in a Lucene-based system which
286 was previously shown to be a robust search engine for biomedical applications¹⁷.
287 Since scientific documents are by and large written in English¹⁸ we have employed
288 standard English analyser and stemmer as parameters of retrieval. We only perform
289 searches on the text of the CS record, disregarding metadata of the full article it was
290 retrieved from.

291

292 Documents are retrieved given a set of keywords to match the text of the CS. A post-
293 processing step after document retrieval is introduced, where we count the number of

294 shared citations between resulting documents. The documents are sorted in
295 descending order firstly by the relevance score of the Lucene system (normalized to
296 one decimal point) and secondly by the number of shared citations. This assures that
297 statements on highly cited papers which are similar within the normalized value of the
298 text-relevance score, are displayed first. The literature search service is available as a
299 web service at <http://finder.sciride.org>. The text-mining of the CS corpus is available
300 through an API which is described on the website.

301

302 **4.4 Information content comparison.**

303

304 We measured how different the CSs are to PR document titles, abstracts and citations.
305 Since a typical search engine operates on the remit of keywords, we have created a
306 textual fingerprint for each CS, title/abstract and full text. Each fingerprint was a set
307 of case-folded, stemmed and stop-word-free normalized words without duplicates.

308

309 For each PR document, we have collected three elements: its title/abstract, full text
310 and a list of CSs referring to it. We have created a textual fingerprint for each
311 title/abstract, full text and each CS, which was supposed to emulate a typical corpus
312 employed by an information retrieval system.

313

314 To produce a fingerprint for a given piece of text, we split it into word tokens using
315 the NLTK toolkit. We case-folded each word, and removed any punctuation (keeping
316 special symbols such as Greek letters). We removed all stop-words (as defined by the
317 NLTK corpus). Finally each word was stemmed so as to minimize mismatches in
318 subtle inflection forms¹⁹. We did not keep word duplicates, thus for each text element
319 (such as title/abstract), this resulted in a non-redundant list of normalized words.

320

321 A typical information retrieval algorithm can be expected to perform such text-
322 normalization operations on a given document. Thus, it is reasonable to assume that if
323 text-normalized fingerprints share many words, the information retrieval algorithm
324 would treat them as contributing similar information and yield similar results.
325 Therefore, the number of different normalized words between CS and PR
326 title/abstract and full text was taken as a measure if CS contribute new information on
327 PR.

328

329 Comparing two fingerprints (e.g. CS versus full-text) consisted of counting how many
330 text-normalized words are found in one fingerprint but not the other.

331

332 **Author Contributions**

333

334 AV and KK designed the experiments wrote the manuscript and prepared the
335 figures. KK wrote the text mining algorithms and created the finder.sciride.org
336 website.

337

338

339 **Additional information**

340

341 The authors declare no competing financial interests related to this work and any
342 material used in this study.

343

344 **References**

345

- 346 1. Neylon, C. & Wu, S. Article-level metrics and the evolution of scientific
347 impact. *PLoS Biology* **7**, (2009).
- 348 2. Beel, J. & Gipp, B. Google Scholar 's Ranking Algorithm: An Introductory
349 Overview. *12th Int. Conf. Sci. Inf.* **1**, 230–241 (2009).
- 350 3. Ostell, J. in *The NCBI Handbook* 1–6 (2002).
- 351 4. Jacso, P. As we may search - Comparison of major features of the Web of
352 Science, Scopus, and Google Scholar citation-based and citation-enhanced
353 databases. *Current Science* **89**, 1537–1547 (2005).
- 354 5. Beck, J. & Sequeira, E. in *NCBI Handbook* 1–17 (2013).
- 355 6. Fernández, J. M., Hoffmann, R. & Valencia, A. IHOP web services. *Nucleic
356 Acids Res.* **35**, (2007).
- 357 7. Chen, H. & Sharp, B. M. Content-rich biological network constructed by
358 mining PubMed abstracts. *BMC Bioinformatics* **5**, 147 (2004).
- 359 8. Fujiwara, T. & Yamamoto, Y. Colil: a database and search service for citation
360 contexts in the life sciences domain. *J. Biomed. Semantics* **6**, 38 (2015).
- 361 9. Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., Jensen, L. J. & Brunak, S.
362 Text mining of 15 million full-text scientific articles. *doi.org* 162099 (2017).
363 doi:10.1101/162099
- 364 10. Hearst, M. A. *et al.* BioText Search Engine: Beyond abstract search.
365 *Bioinformatics* **23**, 2196–2197 (2007).
- 366 11. Xu, S., McCusker, J. & Krauthammer, M. Yale Image Finder (YIF): A new
367 search engine for retrieving biomedical images. *Bioinformatics* **24**, 1968–1970
368 (2008).
- 369 12. Abu-Jbara, A. & Radev, D. Reference scope identification in citing sentences.
370 *12 Proc. 2012 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang.
371 Technol.* 80–90 (2012).
- 372 13. Qazvinian, V. & Radev, D. R. Identifying non-explicit citing sentences for
373 citation-based summarization. in *Proceedings of the 48th Annual Meeting of
374 the Association for Computational Linguistics* 555–564 (2010). doi:Association
375 for Computational Linguistics
- 376 14. Qazvinian, V. & Radev, D. R. Scientific Paper Summarization Using Citation
377 Summary Networks. in *COLING '08 Proceedings of the 22nd International
378 Conference on Computational Linguistics* 689–696 (2008).
379 doi:10.3115/1599081.1599168
- 380 15. Piwowar, H. *et al.* The State of OA: A large-scale analysis of the prevalence

- 381 and impact of Open Access articles. *PeerJ Prepr.* (2017).
382 doi:10.7287/peerj.preprints.3119v1
383 16. Piwowar, H. A., Day, R. S. & Fridsma, D. B. Sharing detailed research data is
384 associated with increased citation rate. *PLoS One* **2**, (2007).
385 17. Yu, H. *et al.* Development, implementation, and a cognitive evaluation of a
386 definitional question answering system for physicians. *J. Biomed. Inform.* **40**,
387 236–251 (2007).
388 18. Ferguson, G., Erez-Llantada, C. & Plo, R. O. English as an international
389 language of scientific publication: a study of attitudes. *World Englishes* **30**, 41–
390 59 (2011).
391 19. Porter, M. F. An algorithm for suffix stripping. *Program* **14**, 130–137 (1980).
392
393
394
395
396
397
398